Lessons Learned from Protein Folding

Silvia Crivelli

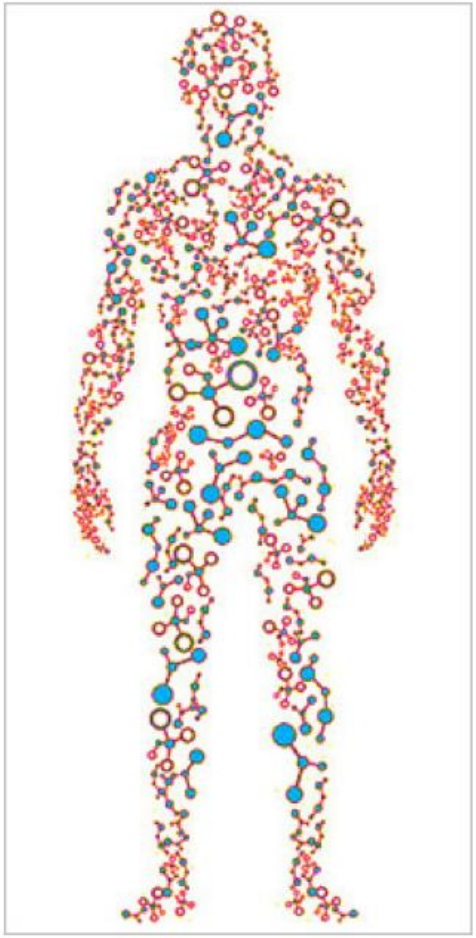
UC Davis and LBNL

What's Protein Folding?

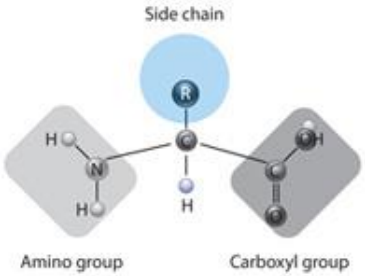
It's to determine how proteins fold themselves into those complex shapes that determine the role they play in life

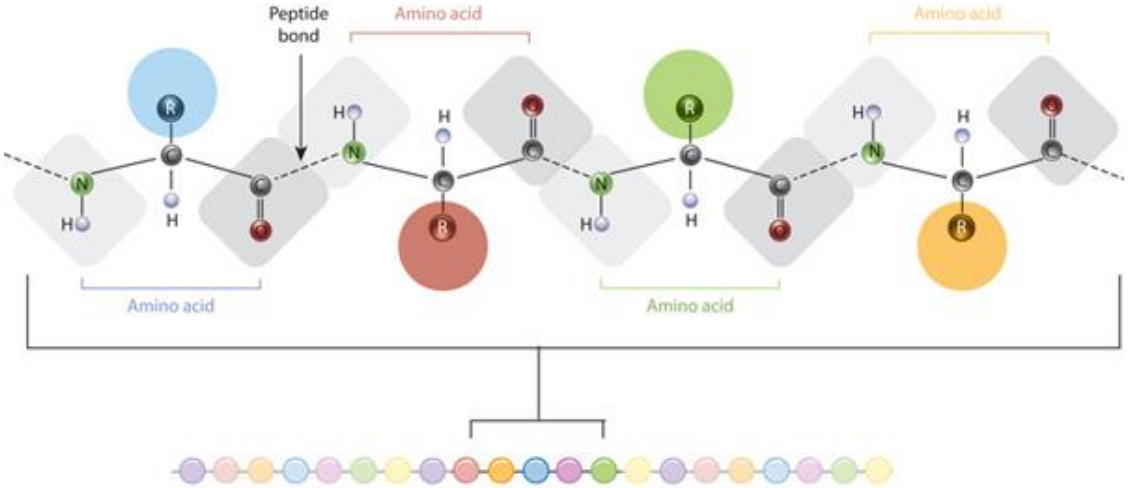
Proteins: the building blocks of life

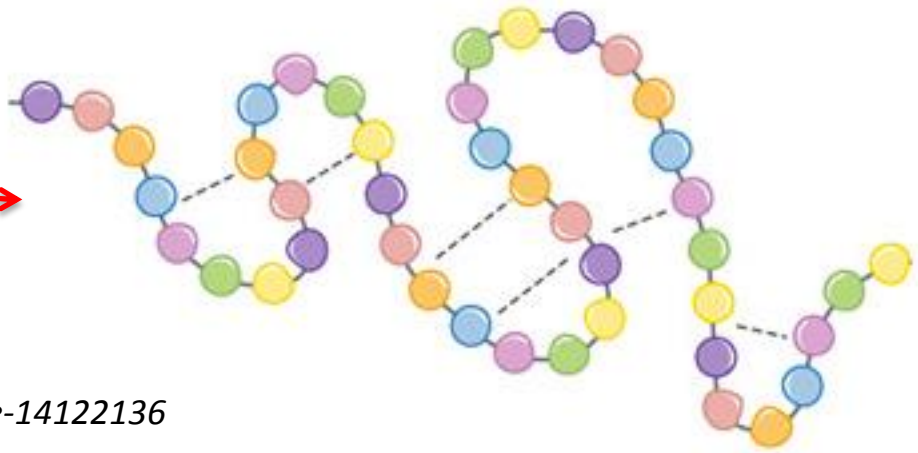
- Associated with most functions in your body
- Associated with diseases
- There are billions or trillions of proteins inside your body
- They are very small



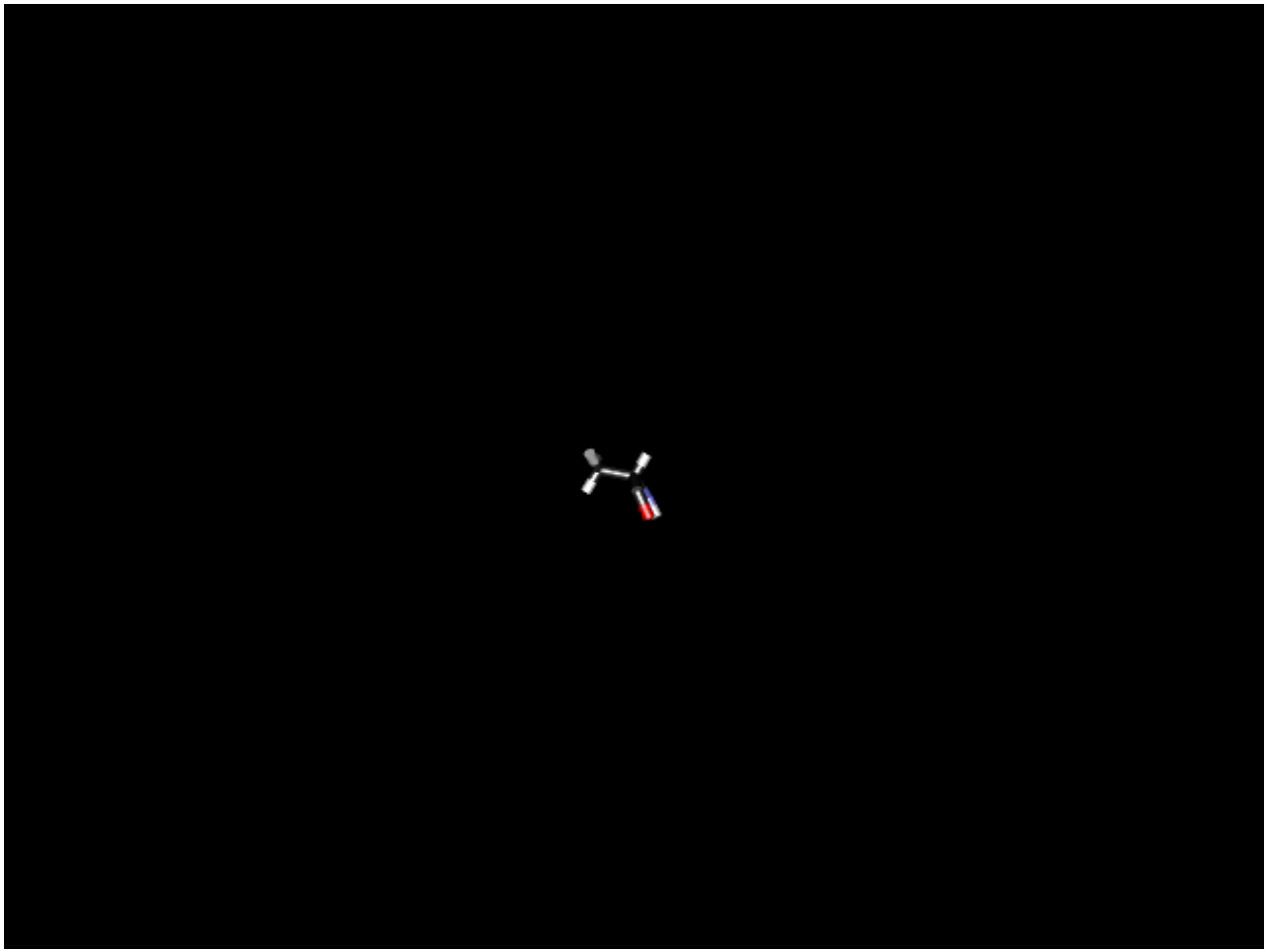
http://www.nih.gov/researchmatters/june2014

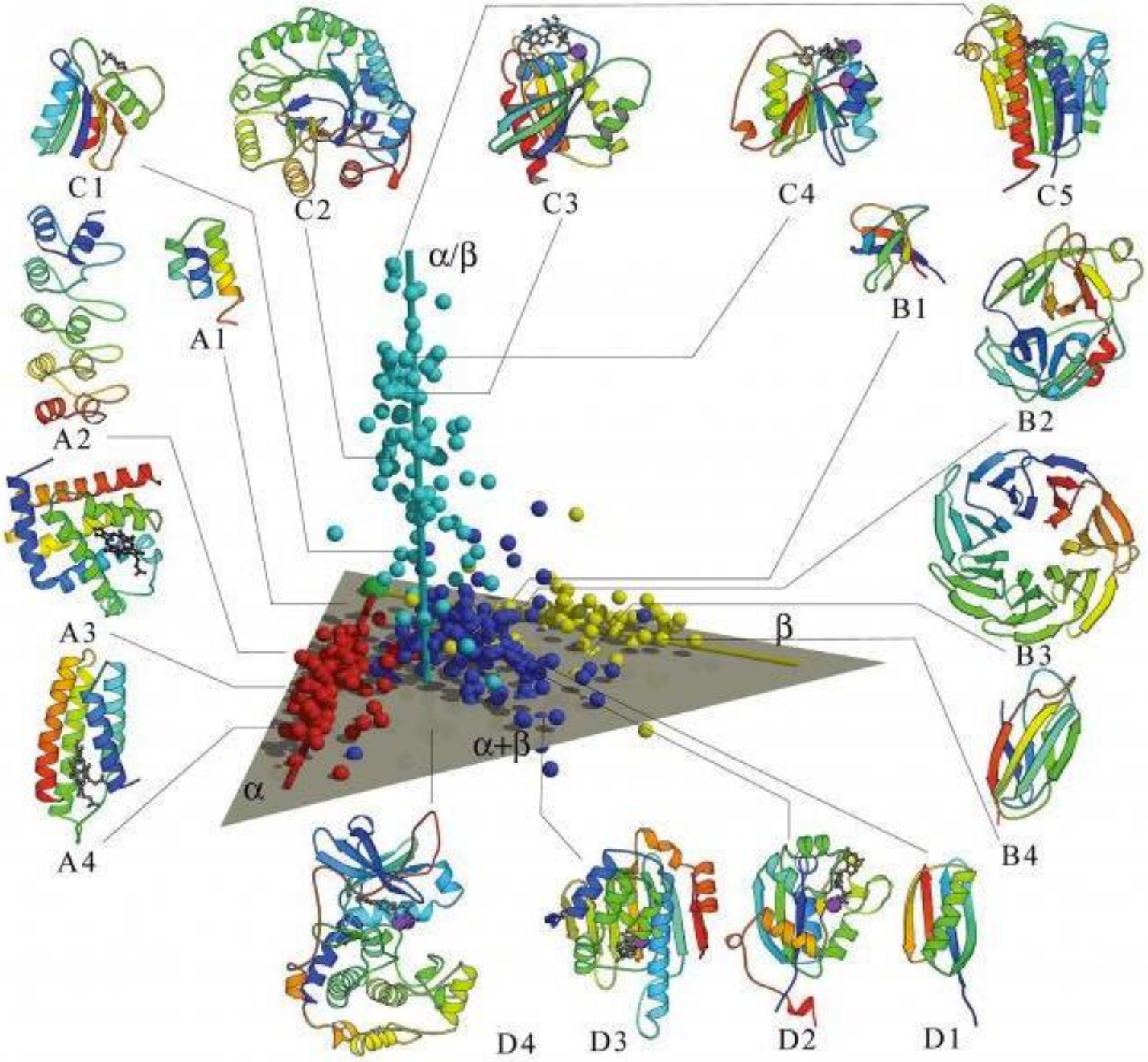




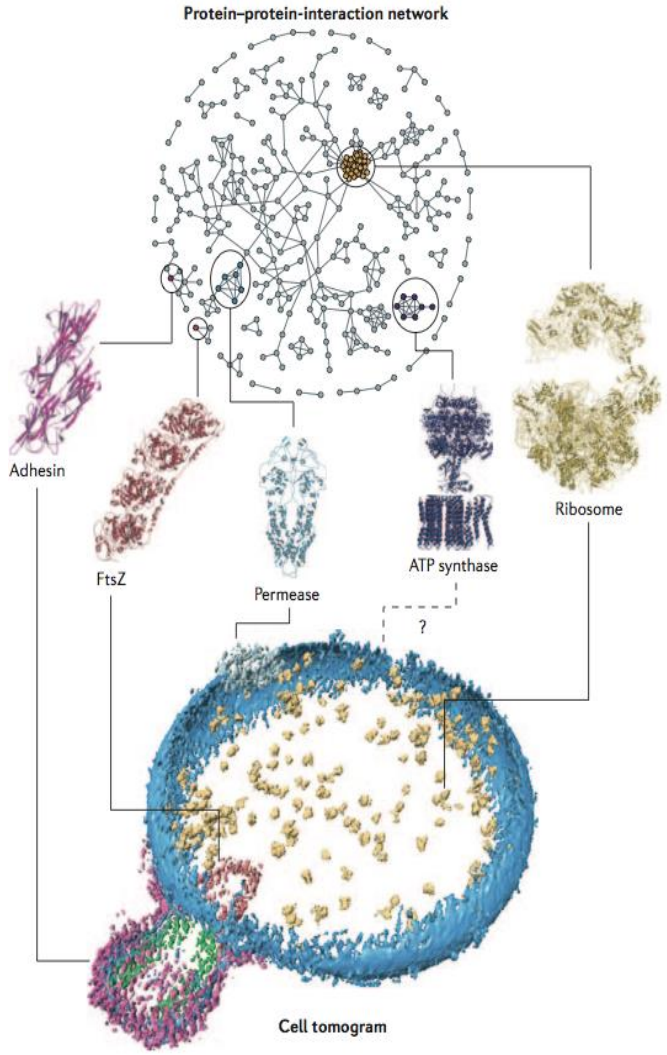


http://www.nature.com/scitable/topicpage/protein-structure-14122136




C1
C2
C3
C4
C5
α/β
A1
A2
A3
A4
B1
B2
B3
B4
β
α+β
α
D4
D3
D2
D1

Molecular Machines



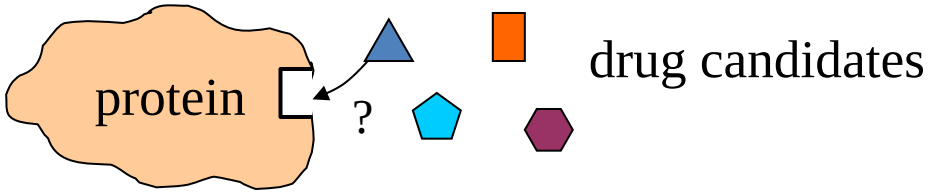
- Create energy,
- pump,
- spin around and
- perform complex functions.
- There are about 20,000 types of machines in our body and 100,000 in living systems.

Molecular Machines: ATP Synthase

Why is Protein Folding so Important?

- ✓ To understand the mechanisms of life
- ✓ To find new drugs to combat disease. Diseases associated with proteins not working properly include:
  - Cystic fibrosis
  - Parkinson's
  - Alzheimer's

  What does the active site look like?



- ✓ To re-engineer defective proteins
- ✓ To design new proteins with desired functions not currently found in nature.
- ✓ Laboratory experiments are expensive:
  - X-ray crystallography
  - NMR

Why is Protein Folding so Hard?

The 3D structure of a protein corresponds to the global minimum of its free energy function[1].

Challenges:

- Formulating an energy function that describes the protein's interactions.
- Large number of local minima.
- Large parameter space (~4800 variables for a 100 amino-acid protein).

[1] Anfinsen et al., PNAC 47, 1961.

CASP: Critical Assessment of protein Structure Prediction

Protein Structure Prediction Center

Menu

- Home
- FORCASP Forum
- PC Login
- PC Registration
- CASP Experiments
  - CASP ROLL
    - Home
    - My CASP ROLL profile
    - Targets
      - Target List
      - Target Submission
  - CASP11 (2014)
  - CASP10 (2012)
  - CASP9 (2010)
  - CASP8 (2008)
  - CASP7 (2006)
  - CASP6 (2004)
  - CASP5 (2002)
  - CASP4 (2000)
  - CASP3 (1998)
  - CASP2 (1996)
  - CASP1 (1994)
- Initiatives
- Data Archive
- Local Services

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been ten previous CASP experiments. The eleventh experiment will start in May 2014. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

CASP1 (1994) | **CASP2 (1996)** | **CASP3 (1998)** | **CASP4 (2000)** | **CASP5 (2002)** | **CASP6 (2004)** | **CASP7 (2006)** | **CASP8 (2008)** | **CASP9 (2010)** | **CASP10 (2012)** | **CASP11 (2014)**

Raw data for the experiments held so far are archived and stored at our data archive.

Starting November 2011, we are opening a new rolling CASP experiment for all-year-round testing of ab initio modeling methods:

CASP ROLL

Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. **CASP proceedings** include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on this web page. The main numerical measures used in evaluations are described in the papers [1], [2]. The latter paper also contains explanations of data handling procedures and guidelines for navigating the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

Message Board

Resuming CASP ROLL

Dear CASPers, Best regards for all of you in the New Year! Hoping that you had good rest after the CASP10 experiment and meeting, we are resuming CASP ROLL with two new targets later this week. ...

Predictors meeting in Gaeta

Dear CASP10 Participants, On the last day of the Meeting we will have our regular Predictors get-together. In advance, I would like to ask you to send in any comments regarding the CASP process in ...

Release of CASP10 results

Dear CASP10 Predictors, We have released results of the CASP10 and CASP ROLL experiments. You can check now interactive results tables and graphs, as well as the parsable data, including the text S ...

Go to "http://www.predictioncenter.org/casp5/Casp5.html"

10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

Target List

Targets expire on specified date at noon (12:00) local time in California (GMT - 7 hours). If information leak occurs after the three weeks since target release, evaluation will be limited to the models submitted within the initial 3 weeks only.

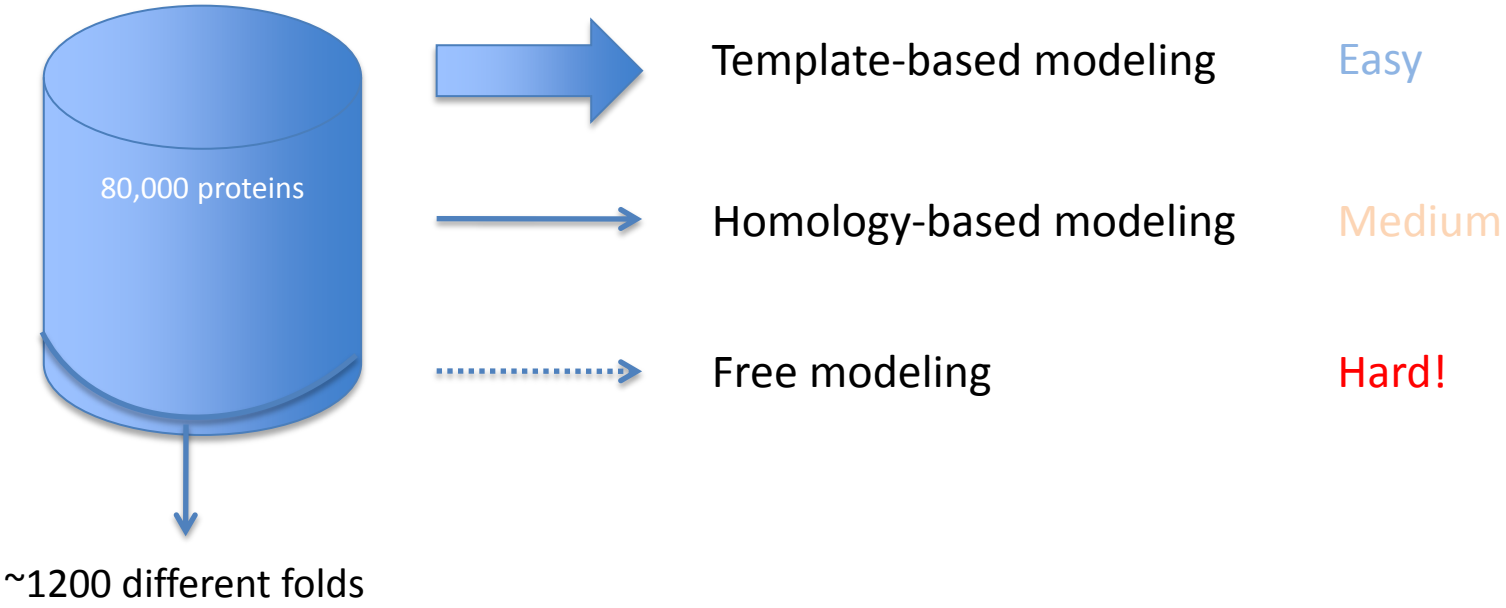
Yellow color - target expires within 48 hours; Orange color - target expires within 24 hours; Red color - target has expired for server TS/DR/RR/FN predictions, but is still open for QA predictions. Special experiment targets are highlighted with the light grey background

All targets | Regular (All groups | Server only) | Refinement | Assisted structure prediction

#	Tar-id	Type	Res	Method	Entry Date	Server Expiration	QA Expiration	Soft 3-week Deadline	Human Expiration	Description
1.	T0644	All groups	166	X-RAY	2012-05-01	2012-05-04	m1: 2012-05-08 m2: 2012-05-10	2012-05-22	2012-05-22	Joint Center for Structural Genomics, GS13193H, PDB code 4fr9
2.	T0646	All groups	93	NMR	2012-05-02	2012-05-05 canceled on 2012-05-03	m1: 2012-05-09 m2: 2012-05-11	2012-05-23 canceled on 2012-05-03	2012-05-23 canceled on 2012-05-03	Northeast Structural Genomics Consortium, OR188, *Canceled - ribbon diagram exposed on the web*
3.	T0649	All groups	210	X-RAY	2012-05-03	2012-05-06	m1: 2012-05-10 m2: 2012-05-12	2012-05-24	2012-05-24	Joint Center for Structural Genomics, GS13209C, PDB code 4f54
4.	T0651	All groups	254	X-RAY	2012-05-04	2012-05-07	m1: 2012-05-11 m2: 2012-05-13	2012-05-25	2012-05-25	Northeast Structural Genomics Consortium, LgR82, PDB code 4f67
5.	T0653	All groups	414	X-RAY	2012-05-07	2012-05-10	m1: 2012-05-14 m2: 2012-05-16	2012-05-28	2012-05-28	Joint Center for Structural Genomics, SP13177B, PDB code 4fs7
6.	T0655	All groups	182	NMR	2012-05-08	2012-05-11	m1: 2012-05-15 m2: 2012-05-17	2012-05-29	2012-05-29	Northeast Structural Genomics Consortium, MiR12, PDB code 2luz

The most successful structure prediction methods are based on assuming that similar sequences lead to similar structures

Protein Data Bank (PDB)



CASP Improvements

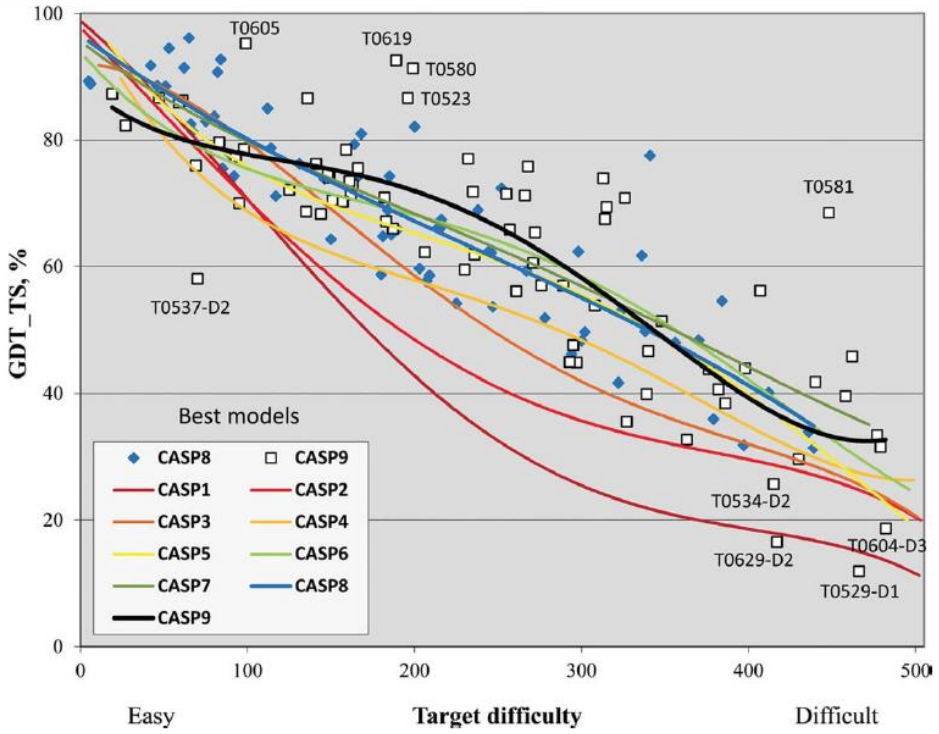
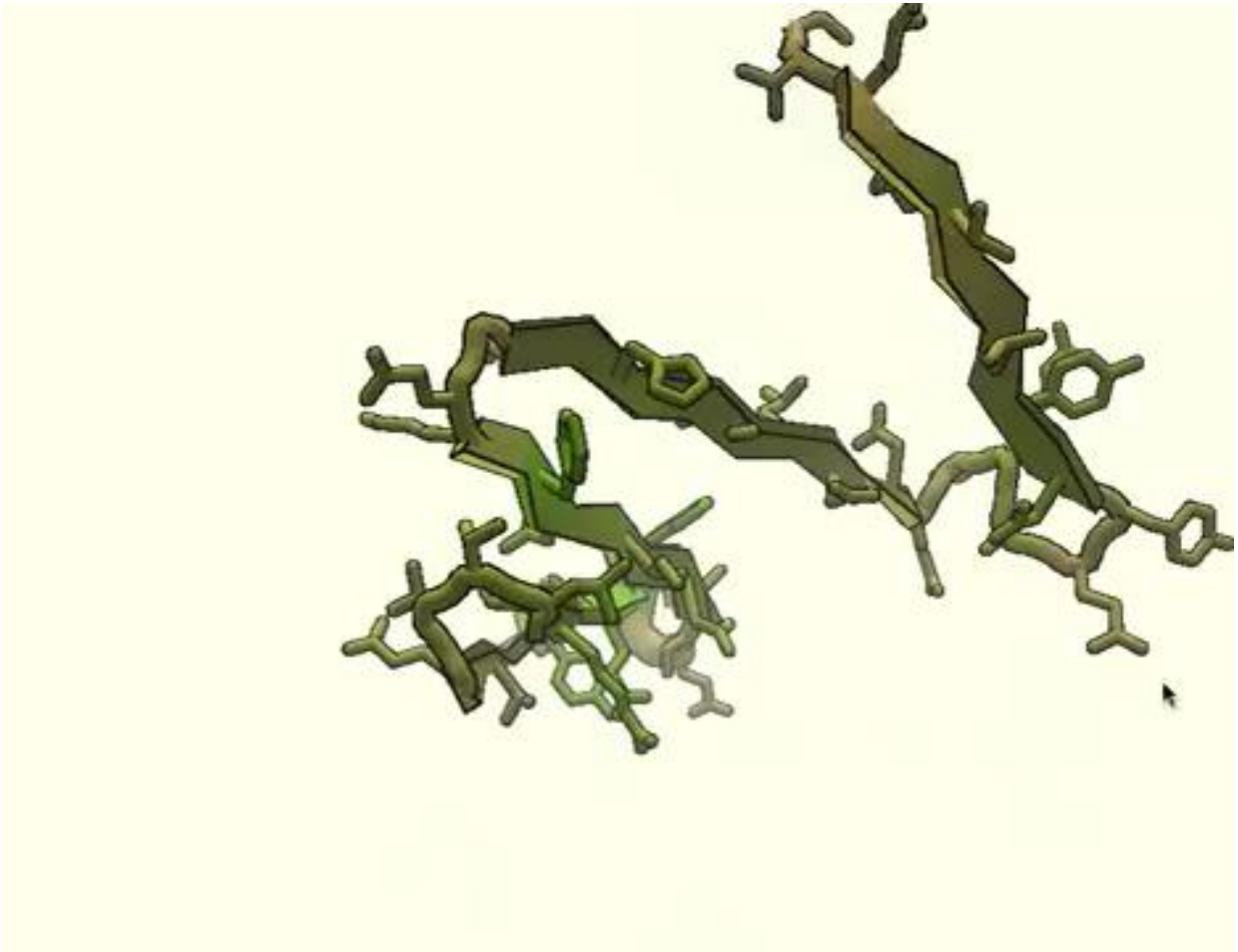


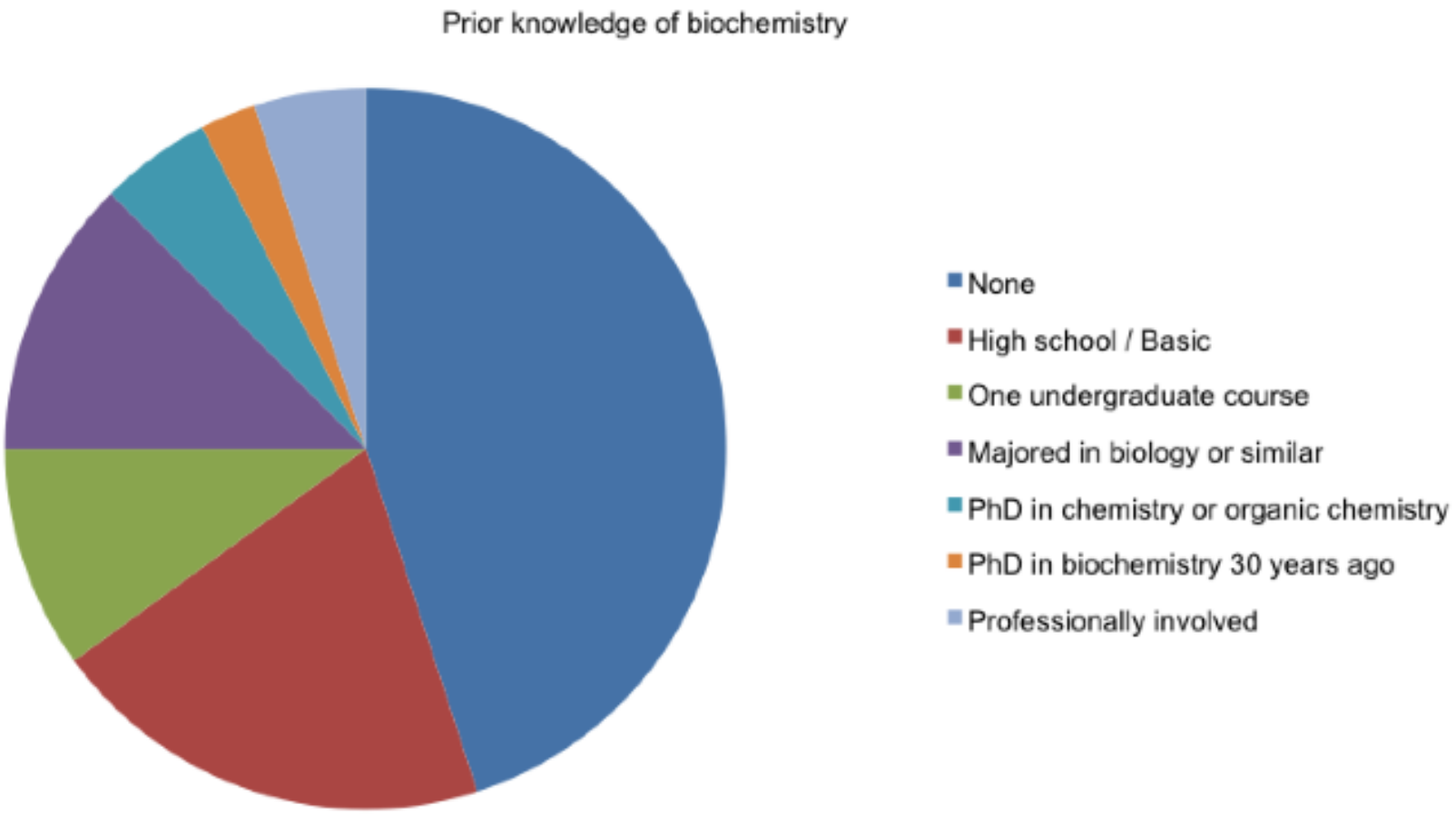
Figure 2

GDT_TS scores of submitted best models for targets in all CASPs, as a function of target difficulty. Each point represents one target. Quartic trend lines show a likely increased accuracy of modeling in the middle range of difficulty in CASP9. Other types of polynomial fit and moving average splines show a similar trend (lines not shown).

The Foldit Game



Foldit Players Demographics



The WeFold Coopetition

- WeFold is an open online coopetition (cooperation + competition) mediated by the WeFold gateway

  http://wefold.nersc.gov.

- It brings together 20 labs worldwide that compete against each other during CASP.
- It provides them the resources to collaborate by contributing different components of their own methods and creating new, hybrid methods.
  - "Each method has a special power"
  - Leverage expertise at a scale not done before

- 19 different structure prediction methods have been developed and are currently being tested during CASP11.

Competition

Lab 1

Lab 2

Lab 3

Coopetition = cooperation + competition

Combined Labs

WeFold1 Results

The WeFold1 teams

- Scored a top prediction for T0740 (one of the hardest targets of CASP10
- Achieved peak performance for TR705 and TR722.

These are very promising results for a first attempt at combining methods without prior preparation or optimization.

What did we learn from our first WeFold experiment?

- We need more labs to increase our chances for success
- We need CASP outsiders: labs with complementary expertise and citizen scientists that can take a look at the problem from a different perspective

WeFold2 as an Educational Tool

- Engage participants outside the CASP community to tackle some specific problems
- Distribute code, data, and educational materials

WeFold2 as an Educational Tool

- My lab is hosting
  - 11 undergrad students (biology, chemistry, computer science, machine learning)
  - 1 grad student (machine learning)
  - 1 high school student
  - 2 visiting faculty

All working under one roof to discuss protein structure prediction and to support the WeFold community

WeFold2 as an Education Tool

- We’re developing protein structure prediction methods
- We’re developing scoring functions using machine learning techniques
- We’re developing a database to support machine learning groups worldwide to develop their own scoring functions
- We’re all learning to work in a multidisciplinary team

Lessons from Protein Folding

- Problem is too hard for one person or one group
- Social-based approaches have been able to advance the field: CASP, Foldit, WeFold
- More importantly, the combination of seasoned researchers, students, and citizen scientists working together increases the chances of success!

Acknowledgments

UCDavis

Prof. Max

Visiting Faculty

Jesse Fox

Rob Hatherill

NERSC/LBNL

Elizabeth Bautista

Colette Flood + WD&E

Shreyas Cholia

Francesca Verdier

Tagrid Samak

Yushu Yao

Students

Nikki Bayar

Stephanie Cabanela

Christopher Cook

Rachel Davis

Ricardo Ferreira

Aubrey Gress

Nathan Lin

Antony Lopez

Kelvin Lu

Jennifer Ogden

Alex Parella

Evan Racah

Rehan Raiyyani

Satinder Singh

And the WeFold community!